Journal of Biomedical Informatics 46 (2013) 830–836

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Defining and measuring completeness of electronic health records for secondary use ☆

Nicole G. Weiskopf [a,*], George Hripcsak [a], Sushmita Swaminathan [b], Chunhua Weng [a]

[a] Department of Biomedical Informatics, Columbia University, New York, NY 10032, United States
[b] Department of Computer Science, Columbia University, New York, NY 10027, United States

ARTICLE INFO

Article history:
Received 30 January 2013
Accepted 22 June 2013
Available online 29 June 2013

Keywords:
Data quality
Electronic health records
Secondary use
Completeness

ABSTRACT

We demonstrate the importance of explicit definitions of electronic health record (EHR) data completeness and how different conceptualizations of completeness may impact findings from EHR-derived datasets. This study has important repercussions for researchers and clinicians engaged in the secondary use of EHR data. We describe four prototypical definitions of EHR completeness: documentation, breadth, density, and predictive completeness. Each definition dictates a different approach to the measurement of completeness. These measures were applied to representative data from NewYork–Presbyterian Hospital's clinical data warehouse. We found that according to any definition, the number of complete records in our clinical database is far lower than the nominal total. The proportion that meets criteria for completeness is heavily dependent on the definition of completeness used, and the different definitions generate different subsets of records. We conclude that the concept of completeness in EHR is contextual. We urge data consumers to be explicit in how they define a complete record and transparent about the limitations of their data.

© 2013 The Authors. Published by Elsevier Inc. All rights reserved.

1. Introduction

With the growing availability of large electronic health record (EHR) databases, clinical researchers are increasingly interested in the secondary use of clinical data [1,2]. While the prospective collection of data is notoriously expensive and time-consuming, the use of an EHR may allow a medical institution to develop a clinical data repository containing extensive records for large numbers of patients, thereby enabling more efficient retrospective research. These data are a promising resource for comparative effectiveness research, outcomes research, epidemiology, drug surveillance, and public health research.

Unfortunately, EHR data are known to suffer from a variety of limitations and quality problems. The presence of incomplete records has been especially well documented [3–6]. The availability of an electronic record for a given patient does not mean that the record contains sufficient information for a given research task.

Data completeness has been explored in some depth. The statistics community has focused extensively on determining in what manner data are missing. Specifically, data may be considered to be missing at random, missing completely at random, or missing not at random [7,8]. Datasets that meet these descriptions require different methods of imputation and inference.

The statistical view of missing or incomplete data, however, is not sufficient for capturing the complexities of EHR data. EHR records are different from research data in their methods of collection, storage, and structure. A clinical record is likely to contain extensive narrative text, redundancies (i.e., the same information is recorded in multiple places within a record), and complex longitudinal information. While traditional research datasets may suffer from some degree of incompleteness, they are unlikely to reflect the broad systematic biases that can be introduced by the clinical care process.

There are several dimensions to EHR data completeness. First, the object of interest can be seen as the patient or as the health care process through which the patient was treated; there is a difference between complete information about the patient versus complete information about the patient's encounters. A patient with no health care encounters and an empty record has a complete record with respect to the health care process, but a blank one with respect to the patient. Furthermore, one can measure completeness at different granularities: the record as a whole or of logical components of the record, each of which may have its own requirements or expectations (e.g., demographic patient information versus the physician thought process) [9,10]. Another

☆ This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike License, which permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.

* Corresponding author. Address: Department of Biomedical Informatics, Columbia University, 622 W 168th Street, VC-5, New York, NY 10032, United States.
E-mail address: nicole.weiskopf@dbmi.columbia.edu (N.G. Weiskopf).

1532-0464/$ - see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.06.010

dimension of completeness emerges from the distinction between intrinsic and extrinsic data requirements. One can imagine defining minimum information requirements necessary to consider a record complete (which could be with respect to either the patient or the health care process), or one can tailor the measurement of completeness to the intended use. Put another way, we can see completeness in terms of intrinsic expectations (i.e., based a priori upon the content) or extrinsic requirements (based upon the use) [11,12].

The EHR data consumers who define these extrinsic requirements will have different data needs, which will in turn dictate different conceptualizations of a complete patient record. Here, Juran's definition of quality becomes valuable: "fitness for use" [12]. It may be that data completeness does not have a simple, objective definition, but is instead task-dependent. Wang and Strong, for example, in their work developing a model of data quality, define completeness as "[t]he extent to which data are of sufficient breadth, depth, and scope for the task at hand" [13]. In other words, whether a dataset is complete or not depends upon that dataset's intended use or desired characteristics. In order to determine the number of complete records available for analysis one must first determine what it means to have a complete patient record. The quality of a dataset can only be assessed once the data quality features of interest have been identified and the concept of data quality itself has been defined [11].

Multiple interpretations of EHR completeness, in turn, may result in different subsets of records that are determined to be complete. The relationships between research task, completeness definition, and completeness findings, however, are rarely made explicit. Hogan and Wagner offer one of the most widely used definitions: "the proportion of observations that are actually recorded in the system" [5]. This definition does not, however, offer specific measures for determining whether a record is complete. Neither does it account for the possibility that completeness may be task-dependent. What proportion of observations should be present? Which observations are desired? Are there any other considerations beyond simple proportion? Furthermore, observations are complex, nested concepts, and it must be determined what level of detail or granularity is needed or expected. In order of increasing detail, one could record a visit that occurred, the diagnoses, all the symptoms, a detailed accounting of the timing of all the symptoms, the clinician's thought process in making a diagnosis, etc.

In the sections below, we enumerate four specific operational and measurable definitions of completeness. These definitions are not exhaustive, but they illustrate the diversity of possible meanings of EHR data completeness. We ran the definitions against our clinical database in order to demonstrate the magnitude of completeness in the database and to illustrate the degree of overlap among the definitions.

2. Materials and methods

Previously, we conducted a systematic review of the literature on EHR data quality in which we identified five dimensions of data quality that are of interest to clinical researchers engaged in the secondary use of EHR data. Completeness was the most commonly assessed dimension of data quality in the set of articles we reviewed [3]. Based upon this exploration of the literature on EHR data quality, consideration of potential EHR data reuse scenarios, and discussion with stakeholders and domain experts, we describe four prototypical definitions of completeness that represent a conceptual model of EHR completeness. Further definitions of completeness are possible and may become apparent as the reuse of EHR data becomes more common and more use cases and user needs are identified.

Fig. 1 presents a visual model of the four definitions of completeness, which are described further in Section 2.1. In this model of EHR data, every potential data point represents some aspect of the patient state at a specific time that may be observed or unobserved as well as recorded or unrecorded. The longitudinal patient course, therefore, can be represented as a series of points over time that may or may not appear in the EHR.

2.1. Definitions

2.1.1. Documentation: a record contains all observations made about a patient

The most basic definition of a complete patient record described in the literature is one where all observations made during a clinical encounter are recorded [5]. This is an objective, task-independent view of completeness that is, in essence, a measure of the fidelity of the documentation process. Assessments of documentation completeness rely upon the presence of a reference standard, which may be drawn from contacting the treating physician [14], observations of the clinical encounter [15], or comparing the EHR data to an alternate trusted data source—often a concurrently maintained paper record [16–19]. Documentation completeness is also relevant to the quality measurements employed by the Centers for Medicare & Medicaid Services [20].

In secondary use cases, however, the data consumer may be uninterested in the documentation process. Instead, completeness is determined according to how well the available data match the specific requirements of the task at hand, meaning that completeness in these situations is more often subjective and task-dependent. While documentation completeness is intrinsic, the following three definitions of completeness are extrinsic and can only be applied once a research task has been identified.

2.1.2. Breadth: a record contains all desired types of data

Some secondary use scenarios require the availability of multiple types of data. EHR-based cohort identification and phenotyping, for example, often utilize some combination of diagnoses, laboratory results, medications, and procedure codes [21–23]. Quality of care and clinician performance assessment also rely upon the presence of multiple data types within the EHR (the relevant data types vary depending upon clinical area) [20,24–27]. More broadly, researchers interested in clinical outcomes may require more than one type of data to properly capture the clinical state of patients [28,29]. In the above cases, therefore, a complete record may be one where a breadth of desired data types is present. It is important to note that the absence of a desired data type in a record does not necessarily indicate a failure in the clinical care process or in the recording process. Rather, it may be that a data

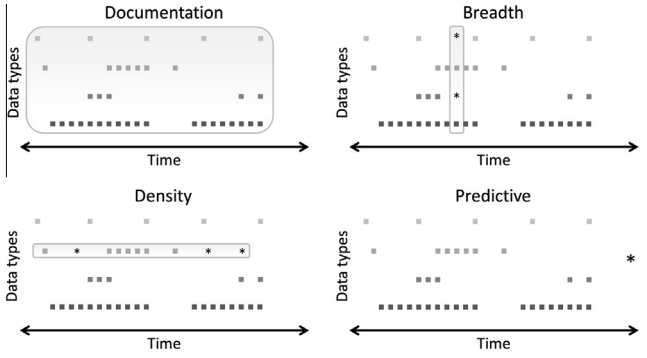

Fig. 1. An EHR completeness model. Each square point denotes an observed and recorded data point, stars are unobserved but desired data points, and the boxes indicate all data points that are required for a given task.

type that is desired for research was not relevant from a clinical standpoint, and therefore was not observed.

2.1.3. Density: a record contains a specified number or frequency of data points over time

In many secondary use scenarios, EHR data consumers require not only a breadth of data types, but also sufficient numbers and density of data points over time [30]. Some of the phenotyping algorithms developed by the eMERGE Network, for example, rely upon the presence of multiple instances of the same laboratory tests, diagnoses, or medications [31], and sometimes specify desired time periods between the recording of these data within the EHR [32,33]. Clinical trial eligibility criteria, which can be compared to patient records to identify relevant cohorts, also contain complex temporal data specifications [34], as do EHR data requests submitted by clinical researchers [35]. Breadth and density can be considered complementary, orthogonal dimensions of completeness. A single point of patient data, for example, has breadth and density of one.

2.1.4. Predictive: a record contains sufficient information to predict a phenomenon of interest

Our final and most complex definition of EHR data completeness arises when one considers that the overall goal of much research is the ability to predict an outcome [13]. It is possible to train various computational models, some of which being more tolerant of missing data than others, using EHR-derived datasets. Researchers may be interested in predicting, amongst other clinical phenomena, disease status and risk [36–38], readmission [39,40], or mortality [41,42]. Depending upon the model employed, data needs may be implicit, rather than explicit. The metric for completeness is performance on the task, rather than counts of data points. The data that are required are those that are sufficient to make a prediction. Therefore, it may that two records with different data profiles are both complete according to this definition.

2.2. Data

NewYork–Presbyterian Hospital (NYPH) is a not-for-profit hospital in New York City consisting of five locations. For the purposes of this research, we included data from: the Milstein Hospital, a tertiary care hospital, and its associated ambulatory areas; Allen Hospital, a community hospital; and Morgan Stanley Children's Hospital. All are in upper Manhattan. These locations and their affiliated offices treat close to 300,000 unique patients per year. The patient population is 56% female, with an average age of 51 years. The population is 32% Hispanic, 10% Asian, 19% Black, and 39% White.

A number of different health information technology systems are in place at NYPH. In this study, we used data from Allscripts's Sunrise Clinical Manager for clinical care, Cerner Millennium for ancillary services, and Eagle Registration for administrative transactions.

2.3. Experiments

Four experiments were designed to demonstrate applications of each of the above definitions to EHR data. A fifth experiment was used to compare the datasets deemed complete according to each of the four definitions. We sampled representative data types for each definition. Specifically, we selected data types that are expected to be present in most EHRs, and which are commonly required in research use cases. These data types include, but are not limited to, admission and discharge information, laboratory results, medication orders, and basic demographic information. Each day a patient was present in the hospital or an affiliated medical office represents an opportunity to observe and record data on the patient state. Each data observation and recording opportunity, in turn, includes multiple data types (e.g., diagnosis, laboratory result, etc.).

EHR completeness can be measured at different levels of granularity. One might examine, for example, the completeness of a full patient record (e.g. each patient represents a potential subject or case), or of specific data types (e.g., lab values are extracted and aggregated across patients). It can also be argued that at any granularity, EHR data never have total completeness. For the purposes of this demonstration, however, we have chosen to measure completeness at the patient record level, and have categorized records as either complete or incomplete according to each definition of completeness. Completeness according to each definition, therefore, is reported in number of patient records that meet the relevant criteria. Rather than provide generalizable completeness findings for EHR data, our goal is to explicitly define and measure completeness from various perspectives and to illustrate the misalignment and intersections among different definitions of completeness.

2.3.1. Documentation

If a complete record must contain all information that was gathered during a clinical encounter—a potential data collection point—a record is incomplete if there was a failure in the recording process. Determining when there was a failure to record data, however, is difficult without a reference standard. NYPH policy dictates that every day that a patient is present in the hospital or one of its affiliated offices, a narrative note should be entered into their record. Therefore, to illustrate this definition, we considered a record without a note on any day that a patient was present for treatment to be incomplete. Inherent in this approach is the assumption that visits are themselves appropriately recorded.

We extracted visit data on all patients in the NYPH clinical data warehouse and determined on which days they were present. Each day was considered to be a potential data collection point. We then identified all days where a patient had a narrative note or report recorded. Every day a patient was present without an associated note or report was said to be a data point that did not meet the definition of documentation completeness.

2.3.2. Breadth

When researchers require a breadth of information about patients, a record is considered complete if certain desired types of information are present. The information required for a record to be deemed complete will vary according to the research task at hand. For this experiment, we chose to look for the presence of five data types frequently found in patient records: laboratory results, medication orders, diagnoses, sex, and date of birth. In this example, a patient with all five data types present would be said to have a complete record. Given the multiplicity of laboratory tests, we also looked specifically at two common laboratory results: blood glucose and hemoglobin measurements. For all patients, we measured the coverage of laboratory results, medication orders, and diagnoses for each day that they were present in the hospital or an affiliated office. The presence of sex and date of birth were assessed once for each patient.

2.3.3. Density

Some research tasks require the availability of multiple data points over time. Moreover, these data points may be required with some degree of regularity or covering a desired period of time. A complete record, therefore, would be one with a desired number of data points over a set period of time, spaced at sufficiently even intervals. For this experiment, we looked at the quantity and temporal distribution of patient visits, medication orders, and

laboratory results. We approached this view of completeness in two ways. First, we looked at the number of clinical data points over the course of a patient record. Second, we applied an adjustment described by Sperrin et al. that accounts for the temporal irregularity of data [30].

$$I = 2/n + \frac{n-2}{n}\left[1 - \sqrt{(n-1)Var\{g_t;\ i = 1,\ldots,n-1\}}\right]$$
$$\text{where } g_t = \frac{x_{i+1} - x_i}{x_n - x_1} \quad (1)$$

I gives the average amount of information provided by each data point by accounting for the variability between those points. In the ideal situation, where all points are evenly spaced, $I = 1$. Multiplying I by n gives the number of effective data points. Sperrin et al. also proposed a linear adjustment that may be used to determine not only how evenly spaced data are, but to what extent a period of interest is covered by those data points. A set of points evenly spaced over a month may give sufficient information about that month, but if the period of interest is a full year, that information becomes insufficient. The adjustment, given a period of interest $[a,b]$, is shown below.

$$I^* = I \times \frac{\min\{b, x_n\} - \max\{a, x_1\}}{b - a} \quad (2)$$

2.3.4. Predictive

One goal of reusing EHR data is to predict something or to find associations. Therefore, a record that contains sufficient information to predict successfully can be considered to be sufficiently complete for the stated purpose. We illustrated the definition for predictive completeness by assessing our ability to predict return visits. Such prediction is important in the context of health care reform, because institutions are striving to reduce readmission rates, and predicting who is likely to return allows institutions to target resources to prevent readmissions. We employed a logistic regression model using type and number of visits, number of medications, and number and value of common laboratory tests as the independent variables and using the presence of a gap of 180 days or more in future visits as the dependent variable.

2.3.5. Comparison of completeness definition results

Further analysis was performed in order to compare records considered to be complete according to the four definitions of completeness. A documentation complete record was one with at least one visit accompanied by a narrative note. Records with breadth completeness were those that included a patient's date of birth, sex, and at least one medication order, laboratory test, and diagnosis. For density, we considered the presence of medication orders and laboratory tests over time, since these data types represent common clinical actions. Temporal resolution was considered down to the second. Sperrin's I was used to calculate the number of effective data points. Finally, we determined the predictive completeness of records using a simplified version of the logistic regression model described in Section 2.3.4. The dependent variable was a gap in each patient record of at least 180 days, and the independent variables were counts of medication orders, laboratory results, and visits in the 3, 6, and 12 months preceding a potential gap.

3. Results

3.1. Documentation completeness

Of the approximately 3.9 million patients with data in the clinical data warehouse, 48.3% have at least one visit recorded where a free-text note or report would be expected. Due to the gradual process of EHR adoption within NYPH, the percentage of missing notes has dropped drastically over the years (Fig. 2). The overall rates of non-missing notes compared to the rates of visits are shown in Fig. 3. Of all the patients with data in the clinical data warehouse, 18.5% have at least one visit with an associated note or report, 7.1% have five or more, and 4% have ten or more. Since 1986, 23.6% of all recorded visits have been accompanied by notes or reports. Over the most recent calendar year, however, the rate of completeness according to this definition has been significantly higher: 98.6% of inpatient visits, 73.8% of outpatient visits, and 95.0% of emergency visits have same day notes or reports recorded.

3.2. Breadth

Of the patients with data in the clinical data warehouse, 29.3% had at least one visit with a recorded laboratory result (20.0% glucose, 23.0% hemoglobin), 12.6% had at least one with a medication order, and 44.5% had at least one with a diagnosis. The vast majority of patient records included basic demographic information: 97.8% had a valid date of birth recorded, and 99.6% had sex recorded.

Fig. 4 shows the rates of visits with associated medications, laboratory tests, and diagnoses, as well as the rates of visits with none, one, two, or all three types of information. Of the patients with records in the clinical data warehouse, 10.4% had at least one visit with all three data types, 26.2% had at least one visit with exactly two, and 33.8% had at least one visit with exactly one.

3.3. Density

Overall, 55.4% of the patients with records in the clinical data warehouse had at least 1 day with a recorded admission event, discharge event, laboratory result, or medication order. Twenty-three point eight percent had at least five, and 15.6% had at least ten. With Sperrin's I applied, 16.6% had at least five, and 10.4% had at least ten. With Sperrin's I and the linear adjustment, these figures dropped even further: 13.6% had at least one, 6.5% had at least five, and 4.4% had at least ten. If the time span of interest is limited to the year in which each patient spent the most days at the hospital, the rates of raw visits and effective visits meeting criteria are lower, but the rates of adjusted visits are higher. Fig. 5 shows the rates of raw, effective, and adjusted counts of days that patients were present in the hospital.

3.4. Predictive

We were able to predict 180-day-or-greater gaps in visits and data with an accuracy of 0.89. The area under the receiver operating characteristic curve was 0.79. The nature of the visits and

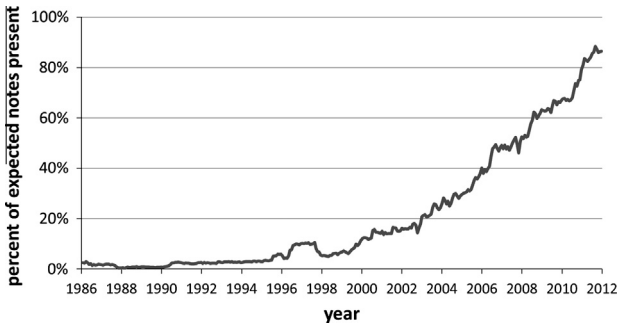

Fig. 2. Documentation completeness improvement over time. The documentation completeness of records has improved as documentation practices have changed and EHR adoption has increased.

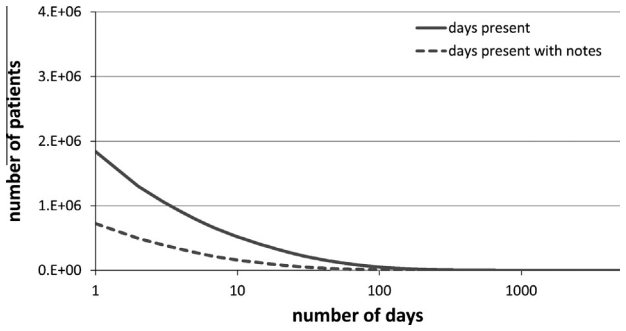

Fig. 3. Documentation completeness of records. Shows the number of patients who have been present in the hospital for a certain number of days, as well as the number of patients whose records have narrative notes or reports associated with a certain number of days that they have been present.

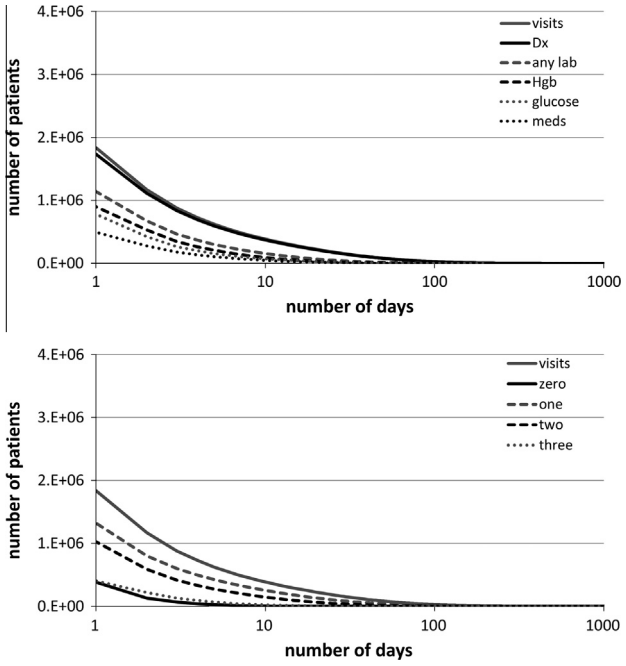

Fig. 4. Breadth completeness of records. The number of patients with laboratory results, medication orders, and diagnoses on the same day as compared to the number of days when they were present in the hospital. Below, the number of patients with zero, one, two, or all of these data types present in their record on the same day.

duration of the records were predictive of gaps. Based upon this conceptualization of completeness, unlike breadth or density completeness, individual cases are predicted either correctly or incorrectly, so there is no sense of an intermediate completeness on an individual case.

3.5. Comparison of completeness definition results

A comparison of the records satisfying the breadth, density, documentation, and predictive definitions of completeness is shown in Fig. 6 [43]. Overall, 55.7% of patients in the CDW have at least one point of clinical data, and 26.9% meet the criteria for at least one definition of completeness. In terms of density, only 11.8% have a complete record when completeness is defined as at least 15 laboratory results or medication orders adjusted for temporal variance. When completeness is defined as a breadth of five data types of interest (date of birth, sex, medication order,

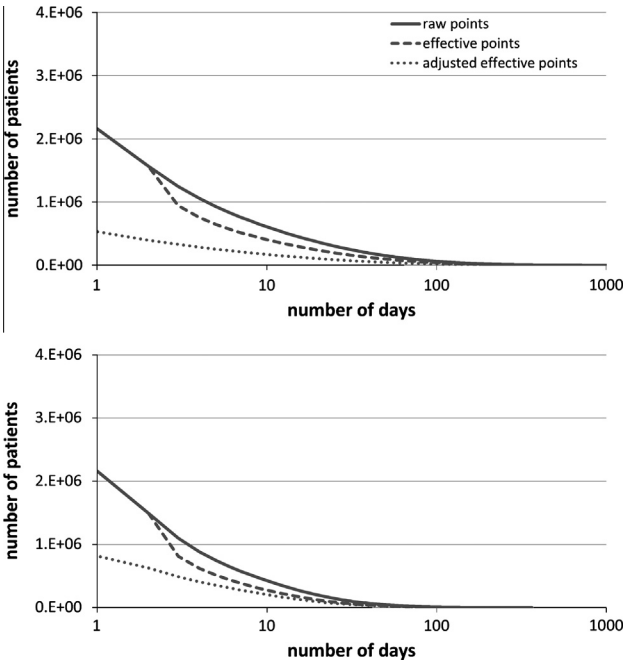

Fig. 5. Density completeness of records. The number of patients with a given number of days with recorded visit events, laboratory results, or medication orders. The raw number of days, the number of days adjusted for variance, and the number of days adjusted for variance and time period are shown.

laboratory test, and diagnosis), 11.4% of patients have complete records. Patients with documentation complete records—meaning they had at least one visit with an associated note—accounted for 18.5% of all patients. Finally, the presence or absence of a gap of 180 days or more could be correctly predicted for 8.4% of patients.

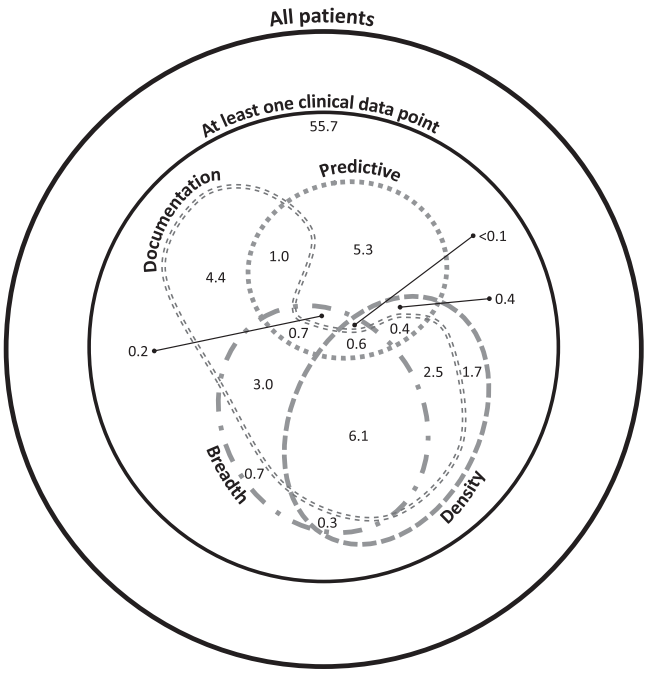

Fig. 6. Comparison of completeness definition results. Subsets of patients with complete records according to the density (medication orders and laboratory tests over time with Sperrin's adjustment), breadth (record includes date of birth, sex, and at least one medication order, laboratory test, and diagnosis), documentation (at least one visit accompanied by a note), and predictive (a gap of 180 days can be correctly predicted) completeness definitions.

Only 0.6% of patient records could be considered complete according to the implementations of all four definitions.

4. Discussion

At the time of this study, the clinical data warehouse contained the electronic records of approximately 3.9 million patients, but the number of records with sufficient information for various analyses is likely much lower. Only about half would be considered complete according to any of the four definitions using the least stringent cutoffs (e.g., at least one data point, at least one visit, or at least one medication or laboratory result). Only about a quarter would be considered complete with more detailed data requirements (e.g., at least one visit with an associated note or laboratory result, at least five visits over the course of a record). When limited not only to complete records, but also to a relevant cohort, the amount of useful information will drop even further. By any definition only a fraction of all the records are complete and suitable for reuse.

Moreover, the number of records in the relevant dataset varies depending upon the definition of completeness being used, which is in turn dependent upon user needs. Someone who is interested in patient care or outcomes over the longitudinal patient record will require very different data from someone looking at a cross-section of a patient population or someone studying the quality of care delivered at a medical institution. These users might identify complete records through, respectively, the density, breadth, and documentation completeness definitions described in this paper. As we have shown, each of these definitions results in a different number of complete records. Before making a determination of how many complete records are available for analysis, therefore, a researcher should first determine and specify what their data needs are, and then select the appropriate definition of completeness and provide it together with the completeness analysis result.

Further complicating the issue of completeness is the fact that not only do different definitions of completeness result in different numbers of useable records, these definitions may also point to different sets of relevant records. One might expect that a record that satisfies one definition of completeness is likely to satisfy another, but this is not necessarily the case. As shown in the comparison of the four definitions of completeness, the resulting sets of useable records share only partial overlap (Fig. 6). In this study, documentation completeness suggests breadth or density completeness, possibly because our method of determining documentation completeness (Section 2.3.1) requires the presence of at least one recorded visit. Predictive completeness, on the other hand, has little overlap with the other three result sets. Although 26.9% of the records in our CDW meet the criteria for at least one of the definitions of completeness, only 0.6% meet the criteria for all four. Therefore, explicitly selecting a relevant definition of EHR completeness is necessary to identify not only how many records are complete, but also which records are complete.

It is important to note that a range of defined completeness is possible and will depend in part upon the complexity of the task for which the data will be used. Taking a trivial example based on the concept of predictive completeness, predicting the patient's age next year requires only the current age, implying most of the patients' records are complete, but predicting the age at which a patient will die is very difficult. Patients with rapidly fatal diseases may be predicted from their diagnoses, but others would be more difficult. Similarly, simple research tasks are likely to require less breadth or density of data than more complex tasks.

There may be analytic ways to address or avoid incompleteness. For example, the algorithm to predict gaps could be used to decide if an individual record is complete. If a patient has a gap and a gap was predicted from preceding data, then perhaps the gap was real; for example, the patient may have been healthy during the period. If, however, the patient has a gap and a gap was not predicted, then perhaps some data are missing. For example, perhaps the patient did have visits but the patient went to a different health provider. Thus the prediction may indicate the likelihood of completeness in the sense of the first definition (i.e., were the data that should have been there present). One could then potentially filter out cases with apparently missing visits.

4.1. Limitations

The rates of complete records identified in this study are not generalizable to other institutions. Differences in populations served, settings, workflows, HIT, and data procedures result in unique data profiles. The definitions of completeness described in this study, however, are not specific to our institution. The idea that information quantity can only be determined following the identification of a relevant definition of EHR completeness and the selection of an appropriate method of measurement is generalizable.

The definitions of completeness described in this study are primarily illustrative and are not exhaustive, as we may have failed to take into account all the needs of potential data consumers. We did not study, for example, the relationship between record completeness and underlying patient status. That is, a healthy patient's record would be expected to look very different from a sick patient's. Further work is needed to more thoroughly and rigorously model the concept of completeness as it relates to the secondary use of EHR data.

The four definitions of completeness described in this study also require further exploration. In the case of predictive completeness, for example, it is unclear how to interpret the result: what level of prediction is sufficient to consider the EHR to be complete? Complicating this is the difficulty distinguishing the cause of low predictive accuracy. It could be because of lack of data, tackling a problem that is hard to solve, or the difficulty of developing an appropriate model.

Finally, completeness is closely tied to other dimensions of data quality. In examining completeness, we made no assumptions regarding the correctness of the data. The fact that data are present does not mean that they are necessarily trustworthy. A full assessment of an EHR-derived dataset prior to reuse should go beyond completeness.

5. Conclusions

We have illustrated that multiple definitions of completeness may be used, that they lead to different degrees of measured completeness for the same dataset, and that the number of complete records in a typical clinical database may be far lower than the nominal total. As researchers and clinicians continue the trend of repurposing EHR data for secondary use, it is important to bear in mind that these clinical data may not satisfy completeness requirements. Completeness, however, is contextual and is determined through an understanding of specific data needs. The number of complete records available for analysis is dependent upon the definition of completeness being used. Each definition results in a different set of complete records. We urge EHR data consumers to be mindful of the potential limitations of a dataset prior to committing to its use, explicit in their choice of completeness definition, and transparent about completeness findings when reporting results.

Funding

This research was supported by National Library of Medicine Grants R01LM006910, R01LM009886, R01LM010815, and

5T15LM007079, as well as Grant UL1TR000040, funded through the National Center for Advancing Translational Sciences.

Acknowledgments

We would like to thank Alla Babina, Yueh-Hsia Chen, and Feng Liu for their assistance in conducting this research. We also thank the anonymous reviewers for providing helpful feedback for an earlier version of this paper.

References

[1] Hersh WR. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. Am J Manag Care 2007;13:277–8.
[2] Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007;14:1–9.
[3] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20:144–51.
[4] Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. BMJ 2003;326:1070.
[5] Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc 1997;4:342–55.
[6] Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev 2010;67:503–27.
[7] Rubin D. Inference and missing data. Biometrika 1976;63:581–92.
[8] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147–77.
[9] Blobel B. Advanced and secure architectural EHR approaches. Int J Med Inform 2006;75:185–90.
[10] Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. AMIA Summits Transl Sci Proc 2010;2010:1–5.
[11] Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc 2002;9:600–11.
[12] Juran JM, Gryna FM. Juran's quality control handbook. 4th ed. New York: McGraw-Hill; 1988.
[13] Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manag Inform Syst 1996;12:5–34.
[14] Lewis JD, Brensinger C. Agreement between GPRD smoking data: a survey of general practitioners and a population-based survey. Pharmacoepidemiol Drug Saf 2004;13:437–41.
[15] Logan JR, Gorman PN, Middleton B. Measuring the quality of medical records: a method for comparing completeness and correctness of clinical encounter data. Proc AMIA Symp 2001:408–12.
[16] Barrie JL, Marsh DR. Quality of data in the Manchester orthopaedic database. BMJ (Clinical research ed.) 1992;304:159–62.
[17] Hohnloser JH, Fischer MR, Konig A, Emmerich B. Data quality in computerized patient records. Analysis of a haematology biopsy report database. Int J Clin Monit Comput 1994;11:233–40.
[18] Ricketts D, Newey M, Patterson M, Hitchin D, Fowler S. Markers of data quality in computer audit: the Manchester Orthopaedic Database. Ann R Coll Surg Engl 1993;75:393–6.
[19] Roukema J, Los RK, Bleeker SE, van Ginneken AM, van der Lei J, Moll HA. Paper versus computer: feasibility of an electronic medical record in general pediatrics. Pediatrics 2006;117:15–21.
[20] Centers for Medicare & Medicaid Services. Quality measures.
[21] Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype–phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010;86:560–72.
[22] Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. Electronic screening improves efficiency in clinical trial recruitment. J Am Med Inform Assoc 2009;16:869–73.
[23] Weng C, Batres C, Borda T, Weiskopf NG, Wilcox AB, Bigger JT, et al. A real-time screening alert improves patient recruitment efficiency. AMIA Annu Symp Proc 2011;2011:1489–98.
[24] Agnew-Blais JC, Coblyn JS, Katz JN, Anderson RJ, Mehta J, Solomon DH. Measuring quality of care for rheumatic diseases using an electronic medical record. Ann Rheum Dis 2009;68:680–4.
[25] Goulet JL, Erdos J, Kancir S, Levin FL, Wright SM, Daniels SM, et al. Measuring performance directly using the veterans health administration electronic medical record: a comparison with external peer review. Med Care 2007;45:73–9.
[26] Jensen RE, Chan KS, Weiner JP, Fowles JB, Neale SM. Implementing electronic health record-based quality measures for developmental screening. Pediatrics 2009;124:e648–54.
[27] Linder JA, Kaleba EO, Kmetik KS. Using electronic health records to measure physician performance for acute conditions in primary care: empirical evaluation of the community-acquired pneumonia clinical quality measure set. Med Care 2009;47:208–16.
[28] Asche C, Said Q, Joish V, Hall CO, Brixner D. Assessment of COPD-related outcomes via a national electronic medical record database. Int J Chron Obstruct Pulm Dis 2008;3:323–6.
[29] Einbinder JS, Rury C, Safran C. Outcomes research using the electronic patient record: Beth Israel Hospital's experience with anticoagulation. In: Proceedings of the annual symposium on computer application [sic] in medical care; 1995. p. 819–23.
[30] Sperrin M, Thew S, Weatherall J, Dixon W, Buchan I. Quantifying the longitudinal value of healthcare record collections for pharmacoepidemiology. AMIA Annu Symp Proc 2011;2011:1318–25.
[31] Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc 2012;19:212–8.
[32] Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. Am J Hum Genet 2011;89:529–42.
[33] Overby C, Pathak J, Haerian K, Perotte A, Murphy S, Bruce K, et al. A collaborative approach to develop an electronic health record phenotyping algorithm for drug-induced liver injury. J Am Med Inform Assoc; 2013, [published online] http//dx.doi.org/10.1136/amiajnl-2013-001930.
[34] Boland MR, Tu SW, Carini S, Sim I, Weng C. EliXR-TIME: a temporal knowledge representation for clinical research eligibility criteria. AMIA Summits Transl Sci Proc 2012;2012:71–80.
[35] Hao T, Rusanov A, Weng C. Extracting and normalizing temporal expressions in clinical data requests from researchers. In: International health informatics conference, Beijing, China; 2013.
[36] Zhao D, Weng C. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. J Biomed Inform 2011;44:859–68.
[37] Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, et al. A predictive model for progression of chronic kidney disease to kidney failure. JAMA 2011;305:1553–9.
[38] Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 2010;48:S106–13.
[39] Cholleti S, Post A, Gao J, Lin X, Bornstein W, Cantrell D, et al. Leveraging derived data elements in data analytic models for understanding and predicting hospital readmissions. AMIA Annu Symp Proc 2012;2012:103–11.
[40] Nijhawan AE, Clark C, Kaplan R, Moore B, Halm EA, Amarasingham R. An electronic medical record-based model to predict 30-day risk of readmission and death among HIV-infected inpatients. J Acquir Immune Defic Syndr 2012;61:349–58.
[41] Wells BJ, Jain A, Arrigain S, Yu C, Rosenkrans Jr WA, Kattan MW. Predicting 6-year mortality risk in patients with type 2 diabetes. Diabetes Care 2008;31:2301–6.
[42] Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Med Care 2013;51:251–8.
[43] Micallef L, Rodgers P. Drawing area-proportional venn-3 diagrams using ellipses. In: 12th Annual grace hopper celebration of women in computing, ACM student research competition and poster session, Baltimore, MD2012.